Mathematical Statistics
Stockholm University
Bachelor Thesis **2022:5**
http://www.math.su.se

Classifying BBC News Articles with Random Forest and eXtreme Gradient Boosting.

Leon Voss Gustavsson*

June 2022

Abstract

With the modern invention and improvement of machine learning methods, the scope of their use is increasing, with new fields of application. One of these fields is natural language processing (NPR). Within NPR one task is text classification. In this thesis we will classify news articles from BBC as belonging to one of the categories sport, politics, entertainment, business or tech. The dataset that we use consists of 2225 observations/articles. Classifying the articles will be done trough analyzing word frequencies from all articles. Furthermore, we will compare different ways of selecting these exact words and study how many words are needed to reach satisfactory results. We will then use Random forest as well as the eXtreme Gradient Boosting (XGBoost) method. In the end we mainly end up with three different factors to evaluate the models by; the number of words used, the computation time, and the prediction accuracy. Despite the reputation of XGBoost, random forest produces somewhat higher prediction accuracy (a fraction of 0.963 correct classification with 10 fold cross validation, compared to 0.957 for XGBoost), and it also takes considerately less time to train. Furthermore, at around 500 words we start to see convergence towards high prediction accuracy for random forest as well as for XGBoost. After having reached such a high degree of prediction accuracy we investigate which words where the most important for classifying an article. Not surprisingly we found that "coach" was meaningful for classifying an article as sport, "shares" for business, and "film" for entertainment and so on. In the end we thus obtained models that where both interpretable and reached high prediction accuracy.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: leon.voss.g@gmail.com. Supervisor: Ola Hössjer, Nils Engler.